Same Content, Different Representations

A Controlled Study for Table QA

Yue Zhang[1] Seiji Maekawa[2] Nikita Bhutani[2]

[1] UT Dallas [2] Megagon Labs

ICLR 2026

Megagon Labs

Motivation

Megagon Labs

Structured Tables

- Fixed schemas, typed columns
- SQL-executable queries
- Precise but brittle

Student	Score	Grade
Alice	92	A
Bob	78	C

Semi-structured Tables

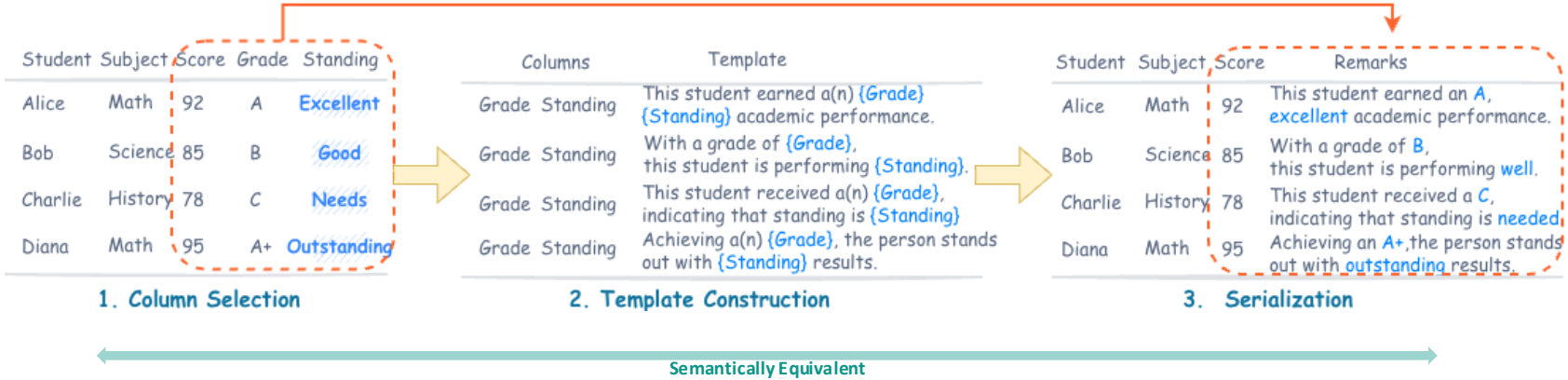
- Merged columns, free text cells
- Irregular schemas
- Flexible but noisy

Student	Remarks
Alice	Earned A, excellent performance
Bob	Scored 78 with C, needs improvement

Key Question: How do different types of table QA methods handle variation in table representation?

Megagon Labs

Approach: RePairTQA Benchmark



Research Questions: RQ1: Structured vs semi-structured? RQ2: Table size? RQ3: Table joins? RQ4: Query complexity? RQ5: Schema quality?

Diagnostic Splits — 4 Dimensions

Table Size

Short (<100) vs Long (≥100 rows)

Table Joins

Single vs Multi-table

Query Complexity

Lookup vs Compositional

Schema Quality

Clean vs Incomplete

Results: Structured vs Semi-structured (RQ1)

Megagon Labs

Method	Family	Structured %	Semi-struct %	Drop %
GPT-4o	LLM	45.4	41.9	**3.4**
Gemini-2.5-flash	LLM	52.1	50.8	**1.3**
Qwen3-235B	LLM	38.2	36.7	**1.5**
LLM-NL2SQL	NL2SQL	**69.1**	38.7	**30.5**
XiYan	NL2SQL	**69.6**	24.1	**45.5**
H-STAR	Hybrid	49.5	47.1	**2.3**
Weaver	Hybrid	62.2	**57.7**	**4.5**

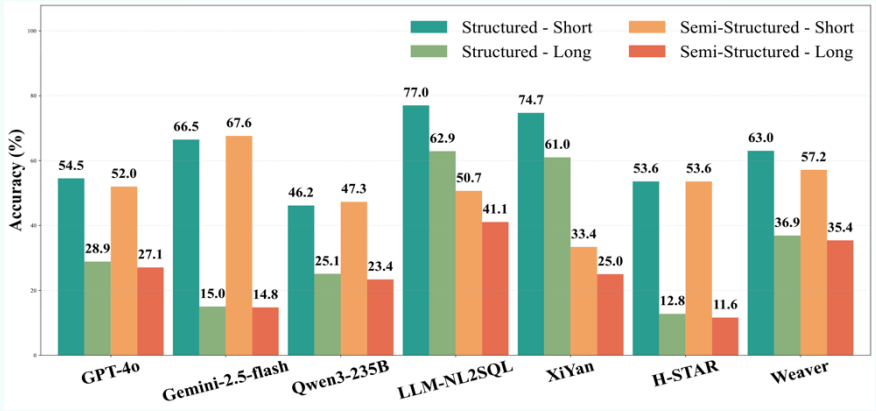
Key Findings

- NL2SQL achieves highest accuracy on structured tables but drops 30–45% on semi-structured
- LLMs are the most robust (only ~1–3% decline) but never reach peak accuracy
- Hybrid methods strike the best balance — Weaver leads on semi-structured (57.7%)

Results: Table Size & Table Joins (RQ2-RQ3)

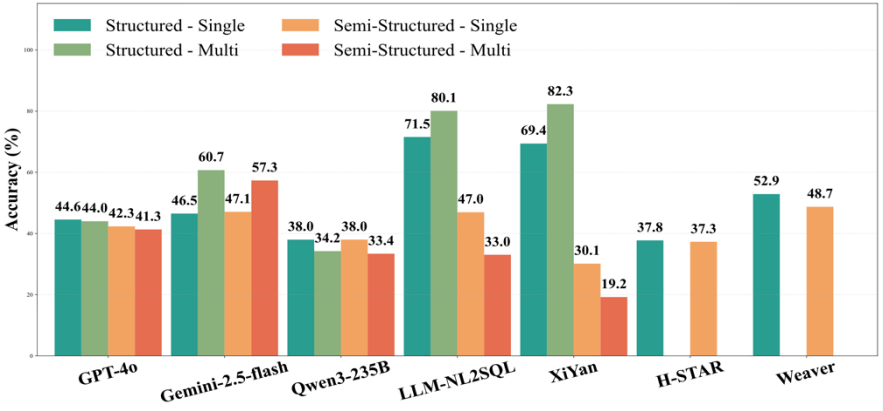
Megagon Labs

RQ2: Table Size



Long tables degrade all methods. LLMs are most sensitive (GPT-4o: 54.5%→28.9%). NL2SQL stays competitive on long structured tables (62.9%). Hybrids offer balanced trade-off: Weaver achieves 57.2% on short semi-structured and 35.4% on long.

RQ3: Table Joins

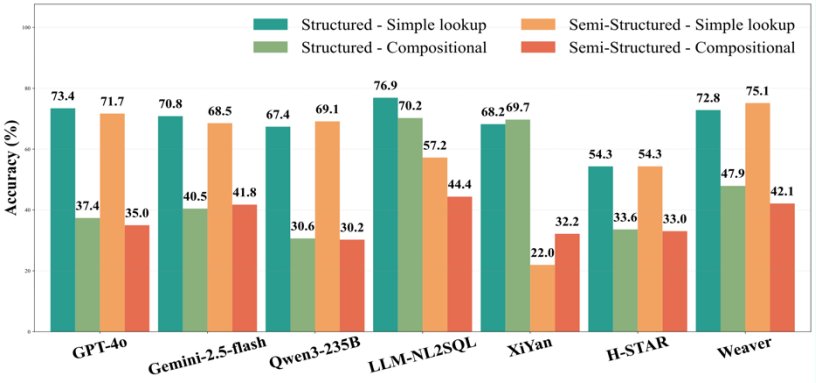


NL2SQL benefits from explicit joins on structured tables (71.5%→82.3%), outperforming GPT-4o by 35+ points. This advantage vanishes on semi-structured inputs where implicit structure hinders grounding. LLMs are largely insensitive to join structure.

Results: Complexity & Schema (RQ4-RQ5)

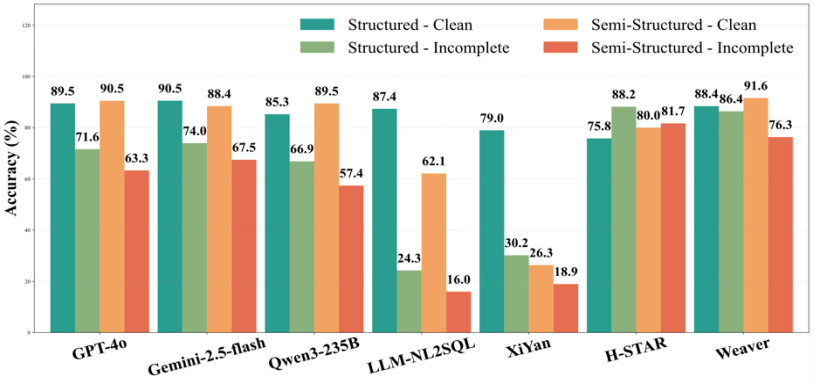
Megagon Labs

RQ4: Query Complexity



All methods drop on compositional queries. LLMs achieve ~70% on simple lookups but struggle with multi-hop reasoning without execution support. Hybrids occupy the middle ground; Weaver sometimes outperforms on semi-structured lookups.

RQ5: Schema Quality



Noisy schemas severely hurt NL2SQL (drop 54.6%). Verbalization often helps LLMs & hybrids by embedding schema cues in natural language. H-STAR mitigates noise via row/column pruning; Weaver remains stable by auto-renaming columns.

Case Study: Structured vs Semi-structured

Structured Table

s_suppkey	s_nationkey	s_comment	s_name	s_address	s_phone	s_acctbal
4248	9	express asymptotes after the …	Supplier#000004248	rst8NdeE8rHKMy38aeq	856-321-5794	7129.18
…	…	…	…	…	…	…

Semi-structured Table

s_suppkey	s_nationkey	s_address	s_phone	description
4248	9	rst8NdeE8rHKMy38aeq	856-321-5794	Supplier#000004248 has an account balance of 7129.18 and is noted for the comment: express …
…	…	…	…	…

Question: Find the supply key of the top ten suppliers with the most account balance, and list the supply key along with the account balance in descending order of account balance.

NL2SQL methods:
correctly ranks the top suppliers by account balance using the explicit **s_acctbal** column. ✓

LLMs:
struggle with compositional reasoning without execution, introducing hallucinated outputs. ✗

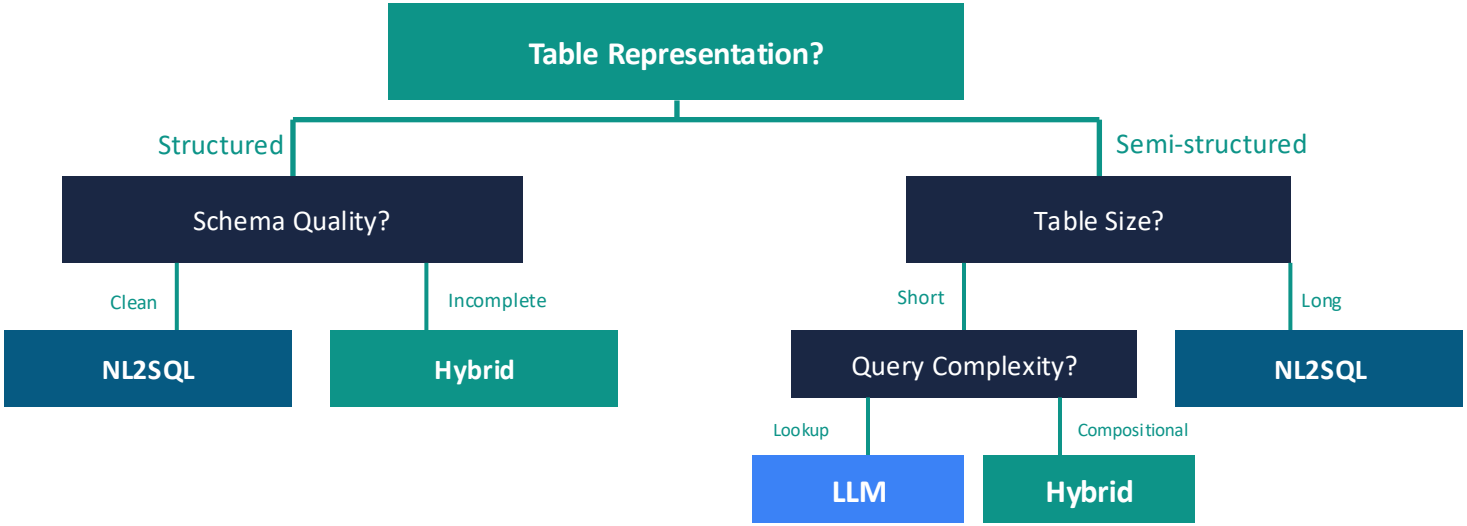
Hybrid Methods:
show mixed outcomes. Weaver producing correct results while H-STAR only outputs reasoning plans. ⚠

NL2SQL methods:
fail on semi-structured tables since account balances are embedded in free-text. ✗

LLMs:
struggle with compositional reasoning without execution, introducing hallucinated outputs. ✗

Hybrid Methods:
break down on semi-structured inputs. Weaver missing several results and H-STAR failing to output answers. ✗

Takeaways & Method Selection



- No single paradigm excels across all conditions
- Representation is a first-order factor in Table QA performance
- Future: representation-aware hybrid systems for diverse real-world formats